A COMPARISON OF ERROR PROBABILITIES FOR VARIOUS DISCRIMINATION PROCEDURES WHEN THE POPULATIONS ARE GAMMA

CHARLES F. HAGER

A COMPARISON OF ERROR PROBABILITIES FOR VARIOUS DISCRIMINATION PROCEDURES WHEN THE POPULATIONS ARE GAMMA

* * * * *

Charles F. Hager

A COMPARISON OF ERROR PROBABILITIES FOR VARIOUS DISCRIMINATION PROCEDURES WHEN THE POPULATIONS ARE GAMMA

by

Charles F. Hager

Lieutenant Commander

United States Navy

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE
WITH MAJOR IN
MATHEMATICS

United States Naval Postgraduate School
Monterey, California

1 9 6 4

A COMPARISON OF ERROR PROBABILITIES FOR VARIOUS DISCRIMINATION PROCEDURES WHEN THE POPULATIONS ARE GAMMA

by

Charles F. Hager

This work is accepted as fulfilling

the thesis requirements for the degree of

MASTER OF SCIENCE

with major in

MATHEMATICS

from the

United States Naval Postgraduate School

TABLE OF CONTENTS

SECTION I

INTRODUCTION

Discrimination analysis has been developed through broad phases in much the same manner as the general history of statistical inference. There have been the Pearsonian phase with the introduction of the coefficient of racial likeness, the Fisherian phase connected with the linear discriminant function, the Neyman-Pearson phase with the introduction of the notions of risk and minimax, and the contemporary Waldian phase. Although the coefficient of racial likeness and generalized distance, proposed by Karl Pearson and P. C. Mahalanobis, respectively are statistics to test the hypothesis of homogeneity, these statistics were the predecessors of discriminatory techniques. It was not until the middle 1930's that R. A. Fisher presented the first clear statement of the problem of discrimination and the first proposed solution to the problem. An excellent survey of the literature on discriminatory analysis and related topics has been compiled by J. L. Hodges in [4].

The general discrimination problem may be classified into three principal types as follows:

(1). <u>A Finite Number of Known Distributions</u> - Let X be a random variable which is known to be distributed according to one of a finite number of

distributions with known density functions, $f_j(x)$, $j = 1, \ldots, m$. On the basis of an observation on X, the problem is to determine which one of the m known distributions is the distribution of X.

(2). <u>Finite Number of Parametric Families of Distributions</u> - Let X be a random variable which is known to have a distribution in one of a finite number of families of distributions. The distributions in the j-th family have density functions, $f_j(x, \varphi_j)$, of known form which depend upon the parameter φ_j which lie in a parameter space Ω_j, $j = 1, \ldots, m$. On the basis of an observation on X, the problem is to determine which one of the j families of distributions is the distribution of X.

(3). <u>Nonparametric</u> - Let T be an individual which is known to belong to one of a finite number of populations, π_j, $j = 1, \ldots, m$. To each individual there corresponds an observable value of a random variable which could be vector-valued. On the basis of a random sample of n_j individuals from population π_j, $j = 1, \ldots, m$, the problem is to decide which one of the m populations contains the individual T as a member.

It may be that the only observation available is the observation on the random variable, X, to be classified, but usually, there are, in addition to the

observation to be classified, other observations available which can be used to estimate the distributions to which X is to be assigned.

The nonparametric type of discrimination problem has received least attention to date. In [2], Hodges and Fix have considered the problem of nonparametric classification in the case of two populations and have developed procedures which were shown to have asymptotic optimum properties for large samples. In [3], Hodges and Fix compared several of these nonparametric procedures against the linear discriminant function when the two populations are normal with equal covariance matrices. The linear discriminant function is a widely employed classification procedure, and therefore, it is of interest to determine the performance of this procedure when the populations are not gaussian. In [1], Thomas E. Eaton compared one of the nonparametric procedures proposed in [2] against the linear discriminant function when the two populations were exponential. The basis of comparison in both [1] and [3] was the probability of misclassification. This thesis is a continuation of the research started in [1].

Section II will summarize the procedures and results of [3] as all of the procedures used in this paper are analogous. Section III provides a complete comparison of the probabilities of misclassification

of a nonparametric procedure against the linear discriminant function when the two populations are exponential. Section III also includes a limited tabulation of the probabilities of misclassification for the linear discriminant function when the two populations are gamma and one of the parameters has its domain restricted to the positive integers. Due to time limitation, it was not possible to determine a satisfactory computational formula to compute the probabilities of misclassification for the nonparametric procedure when the two populations are gamma. Section IV contains conclusions and recommendations based on the results obtained in Section III.

I am indebted to Professor J. R. Borsting for his encouragement and most capable guidance and advise while acting as faculty advisor, and wish to thank Professor R. R. Read for his valuable assistance and advice as second reader. Also, I wish to thank and acknowledge Mrs. Patricia Johnson for programming the procedures developed in Section III of this thesis.

SECTION II

PERFORMANCE OF THE LINEAR DISCRIMINANT FUNCTION AND A NONPARAMETRIC DISCRIMINATOR WHEN THE TWO POPULATIONS HAVE NORMAL DISTRIBUTIONS WITH EQUAL COVARIANCE MATRICES

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be samples from the p-variate distributions F and G, respectively, and let Z be an observation known to be from either F or from G; on what basis is it decided to which population Z belongs? When F and G are p-variate normal distributions with equal covariance matrices, the linear discriminant function is known to be an appropriate procedure. But what is a reasonable procedure when the parametric forms of F and G are not known?

In [2], Hodges and Fix suggest, as an intuitive approach, the following nonparametric procedure: Define in p-dimensional space a notion of distance which will permit a ranking of the 2n observations according to their nearness to Z. Then select an odd integer, k, and assign Z to that distribution from which came the majority of the k nearest observations. Several classes of these nonparametric discriminators are shown to have asymptotically optimum performance in the sense that the probabilities of misclassification,

$P_1 = P[Z \text{ is assigned to } G | Z \text{ came from } F]$

$P_2 = P[Z \text{ is assigned to } F | Z \text{ came from } G]$

tend, as n tends to infinity, to the theoretical

minimum values if F and G were completely known. Since it would not be reasonable to employ a nonparametric procedure solely on the basis of asymptotic properties and applicational simplicity, an investigation is made in [3] to determine how much discriminating power is lost through the use of a nonparametric discriminator when samples are small. To this end, Hodges and Fix assume that F and G are normal with equal covariance matrices so that the linear discriminant function is appropriate. Then a comparison of the probabilities of misclassification, P_1 and P_2, which result when the linear discriminant function is employed with the corresponding probabilities P_1 and P_2 obtained when an alternate nonparametric discrimination procedure is used, indicates how much discriminating power is lost when sample sizes are small. The remainder of this Section is devoted to summarizing some of the procedures and results of [3].

The principal distance function compared with the linear discriminant function is

$$(1). \quad \Delta\,(x,z) = \operatorname*{Max}_{i=1}^{p} |x_i - z_i|$$

although Δ is just one of a large class of distance functions, anyone of which could be used. This fact is mentioned since the probabilities of error, P_1 and P_2, depend very heavily on the distance function

employed. Also, a great part of the computations are made using $k = 1$, that is, assign Z to the population F or G from which came the individual of the pooled samples which most closely resembles Z. This case will be denoted the rule of the "nearest neighbor."

By considering linear transformations on the observation space, the problem can be reduced considerably since it is always possible by such transformations to ensure F and G will have the identity covariance matrix. Thus, the p transformed measurements have unit variance and are independent in each population. Also, it is possible by such transformations to place the expectation vector of F at the origin and the expectation vector of G on the positive first axis. In performing such linear transformations, the probabilities of misclassification, P_1 and P_2, are unchanged for both the nonparametric procedure and linear discriminant function. Thus without loss of generality, it is sufficient to consider the transformed populations with the two parameters, p and λ, where

$$\lambda = E(\text{first coordinate of } Y)$$

$$= \text{distance between the means of the transformed populations.}$$

Furthermore, from the symmetry of the problem it is evident that $P_1 = P_2$ for both procedures; consequently, it is sufficient to compute P_1, that is, assume Z is distributed according to F.

For the univariate case, $p = 1$, the linear discriminant function is greatly simplified since no matrix computation occurs. The procedure consists simply of computing the arithmetic mean of the sample means,

$$\frac{\bar{X} + \bar{Y}}{2},$$

and assigning Z to that population whose sample mean lies on the side of $(\bar{X} + \bar{Y})/2$ as does Z itself. The probabilities of misclassification are readily computed by introducing two new variables which are functions of $\bar{X}$, $\bar{Y}$, and Z. The exact procedure is outlined in [3], but not included in this summary since the subsequent investigation does not depend upon this technique. Table 1 provides a tabulation of values of $P_1 = P_2$ for various values of n and λ. All tables in this section have been reproduced from [3].

For $p = 1$, the distance function Δ corresponds to ordinary Euclidean distance and the nonparametric procedure using the "rule of the nearest neighbor," $k = 1$, consists of assigning Z to that population from which came the sample individual nearest to Z. The probability, P_1, that the nearest neighbor to Z is one of the Y's, given that Z is distributed as X, is readily computed using the following technique. Define $P_1(z)$ to be the conditional probability that the nearest of the 2n sample observations to Z is a Y, given that

$Z = z$. Hence,

$$(2). \qquad P_1 = E\,[P_1(z)] = \int_{-\infty}^{\infty} f(z)\, P_1(z)dz$$

where f is the density function corresponding to F. Continuing exactly as in [3], it remains only to calculate $P_1(z)$. The event, "the nearest sample value to z is a Y" may be classified into n exclusive events, "the nearest sample value to z is Y_i," $i = 1, 2, \ldots, n$ where the $|Y_i - z|$ are independent identically distributed random variables. By defining

$$H_z(\delta) = P(|X - z| < \delta)$$

and

$$K_z(\delta) = P(|Y - z| < \delta),$$

it is readily shown that the density function for the minimum of the $|Y_i - z|$, $i = 1, 2, \ldots, n$ is

$$n\,[1 - K_z(\delta)]^{n-1}\, dK_z(\delta)$$

and that $P_1(z)$ can be computed by the formula

$$(3). \qquad P_1(z) = n \int_0^{\infty} [1-H_z(\delta)]^n\,[1-K_z(\delta)]^{n-1}\, dK_z(\delta)$$

Formulae (2) and (3) form the basis for all the computations for the "nearest neighbor rule" for any p. Tables 2 and 2A provide a tabulation of $P_1 = P_2$ for the nonparametric discriminator, $k = 1$, for various values of n and λ.

It was shown in [3] that for large n,

$$(4).\quad P_1 \cong E\left[\frac{g(z)}{f(z)+g(z)}\right] = \int_{-\infty}^{\infty} \frac{g(z)f(z)dz}{f(z)+g(z)}$$

The above formula was obtained from an expansion of formula (3) and is quite general. An application of Schwartz's inequality to formula (4), shows the integral can not exceed $\frac{1}{2}$.

Also investigated in [3] are the following additional cases:

(i) A nonparametric procedure using Δ as a distance function with $k > 2$ for the univariate and bivariate normal distributions.

(ii) A nonparametric procedure using Δ as a distance function with $k = 1$, $n = 1$, and $p \geq 2$.

(iii) The effect of other distance functions on the probabilities of misclassification for the bivariate normal distribution.

Due to laborious computations, the investigation of several of the above cases was quite limited, but the results that were obtained indicate that the nonparametric procedures gave "reasonable" error probabilities in cases (i) and (ii). Although for the bivariate normal distribution, different distance functions produced vastly different error probabilities in some instances.

TABLE 1

PROBABILITY OF ERROR, LINEAR DISCRIMINANT FUNCTION, UNIVARIATE NORMAL DISTRIBUTIONS

n	λ =1	λ =2	λ =3
1	.4175	.2532	.1235
2	.3821	.1999	.0910
3	.3611	.1819	.0826
4	.3472	.1744	.0787
5	.3376	.1707	.0763
10	.3175	.1646	.0716
20	.3110	.1616	.0692
50	.3094	.1599	.0678
∞	.3085	.1587	.0668

n = size of sample taken from each population

λ = distance between the means of the two populations

Probability of error = P (Z is assigned to G | Z came from F)

= P (Z is assigned to F | Z came from [illegible]

TABLE 2

PROBABILITY OF ERROR, NONPARAMETRIC DISCRIMINATOR WITH k=1, UNIVARIATE NORMAL DISTRIBUTION

n	λ =1	λ =2	λ =3
1	.4175	.2532	.1235
2	.4086	.2364	.1084
3	.4052	.2307	.1036
4	.4032	.2280	.1014

TABLE 2-A

APPROXIMATE PROBABILITY OF ERROR, NONPARAMETRIC DISCRIMINATOR WITH k=1, UNIVARIATE NORMAL DISTRIBUTION

n	λ =1	λ =2	λ =3
4	.403	.226	.102
5	.401	.225	.100
10	.399	.223	.098
20	.398	.224	.098
50	.398	.225	.098
00	.398	.225	.098

n = size of sample from each population

λ = distance between the means of the two populations

Probability of error = P(Z is assigned to G | Z came from F)

= P(Z is assigned to F | Z came from G)

SECTION III

PERFORMANCE OF THE LINEAR DISCRIMINANT FUNCTION AND THE "RULE OF NEAREST NEIGHBOR" WHEN THE TWO POPULATIONS HAVE GAMMA DISTRIBUTIONS

The validity of the linear discriminant function when the data is obviously not normal has been of great concern to many users and also potential users of this discrimination procedure. In [1], T. E. Eaton investigated the performance of the linear discriminant function and a nonparametric procedure for sample size one and two when the univariate distributions, F and G, are assumed to be exponential with parameters λ and μ respectively. This investigation was performed by computing the probabilities of misclassification. The results of this study showed that both the linear discriminant function and nonparametric discriminator using Δ as a distance function and "the rule of nearest neighbor" can give high probabilities of misclassification for sample size one and two. In this section, the investigation started in [1] is continued in order to provide a limited indication of how much discriminating power the linear discriminant function and "rule of nearest neighbor" have when the populations are not normal.

The scope of the present study is an investigation of the probabilities of misclassification,

$P_1 = P$ [Z is assigned to $G|Z$ came from F]

$P_2 = P$ [Z is assigned to $F|Z$ came from G] ,

for the two population classification problem when the following two procedures are employed:

(i) The nonparametric procedure employing Δ as a distance function and using the "rule of the nearest neighbor," $k = 1$, when F and G are exponentially distributed with parameters λ and μ, respectively, and $\lambda = c\mu$ where c is greater than zero.

(ii) The linear discriminant function when F and G have gamma distributions with parameters (r, λ) and (r, μ) respectively, where r is a positive integer, and, as above, $\lambda = c\mu$ where c is greater than zero.

The density functions of F and G will be denoted by $f(x;r,c\mu)$ and $g(y;r,\mu)$ respectively where

$$(5). \quad f(x;r,c\mu) = \frac{(c\mu)^r \, x^{r-1}}{\Gamma(r)} \quad \exp(-c\mu x)$$

and

$$(6). \quad g(y;r,\mu) = \frac{\mu^r \, y^{r-1}}{\Gamma(r)} \quad \exp(-\mu y)$$

Obviously, when $r = 1$ in formula (5) and (6) above, $f(x;\ 1,\ c\mu)$ and $g(y;\ 1,\ \mu)$ are exponential.

A computation formula for the error probabilities, P_1 and P_2, will be developed first for the "rule of nearest neighbor," procedure (i) above. This procedure consists of assigning Z to that population from which came the sample individual nearest to Z.

Assuming equal samples, say n, are available from each population, it is observed that the folowing relation,

(7). $P_1(n,c) = P_2(n,1/c)$

exists between the error probabilities when F and G have gamma distributions with density functions defined by formulas (5) and (6); hence, this relationship exists when F and G are exponential. Using exactly the same technique as was outlined in Section II, it is observed that if Z = z, and $\delta \geq 0$, then

$$H_z(\delta) = P(|X-z| < \delta) = \begin{cases} \int_0^{z+\delta} f(x;r,c\mu)dx, & \text{if } \delta \geq z \\ \int_{z-\delta}^{z+\delta} f(x;r,c\mu)dx, & \text{if } \delta \leq z \end{cases}$$

$$K_z(\delta) = P(|Y-z| < \delta) = \begin{cases} \int_0^{z+\delta} g(y;r,\mu)dy, & \text{if } \delta \geq z \\ \int_{z-\delta}^{z+\delta} g(y;r,\mu)dy, & \text{if } \delta \leq z \end{cases}$$

It follows from formulas (2) and (3) of Section II that

$$P_1(n,c) = n\int_0^{\infty} f(z;r,c\mu)dz \int_z^{\infty} [1-H_z(\delta)]^n\ [1-K_z(\delta)]^{n-1}\ dK_z(\delta)$$
$$+ n\int_0^{\infty} f(z;r,c\mu)dz \int_0^{z} [1-H_z(\delta)]^n\ [1-K_z(\delta)]^{n-1}\ dK_z(\delta).$$

Hence, by the simple change of variables, $\delta' = c\delta$, $z' = cz$, $y' = cy$ and $x' = cx$, it follows that

$$P_1(n,c) = n\int_0^\infty f(z;r,\mu)dz \int_z^\infty [1-H_z(\delta)]^n\,[1-K_z(\delta)]^{n-1}\,dK_z(\delta)$$

$$+\; n\int_0^\infty f(z;r,\mu)dz \int_0^z [1-H_z(\delta)]^n\,[1-K_z(\delta)]^{n-1}\,dK_z(\delta)$$

$$= P_2(n,1/c)$$

Unfortunately, it was only possible to determine a suitable computational formula for $P_1(n,c)$ when F and G are assumed to have exponential distributions. A preliminary survey indicated that a large computational program would be required if F and G are assumed to have the gamma distributions defined at the beginning of this section.

When F and G are assumed to be exponential, a suitable computation formula for $P_1(n,c)$ is obtained as follows: First, let $z' = \mu z$, $\delta' = \mu\delta$, integrate and combine terms to obtain

$$P_1(n,c) = \frac{c}{(c+1)(2nc+2n+c)} +$$

$$2nc\int_0^\infty \exp(-cz-z)dz \int_0^z [1-2\exp(-cz)\sinh c\delta]^n \cdot$$

$$[1-2\exp(-z)\sinh\delta]^n \cosh\delta\; d\delta$$

Then by interchanging the order of integration and expanding both $[1-2\exp(-cz)\sinh c\delta]^n$ and $[1-2\exp(-z)\sinh\delta]^{n-1}$ into binomial series, it can be shown that

$$P_1(n,c) = \frac{c}{(c+1)(2nc+2n+c)}$$

$$+ \; nc \sum_{k=0}^{n} \binom{n}{k} \sum_{j=0}^{n-1} \binom{n-1}{j} \frac{1}{(ck+j+c+1)} \cdot$$

$$\sum_{i=0}^{k} \binom{k}{i} (-1)^i \sum_{p=0}^{j} \binom{j}{p} \quad (-1)^p \left[\frac{1}{F_{k,j,i,p}} + \frac{1}{F_{k,j,i,p}+2}\right]$$

where $F_{k,j,i,p} = (2ck+2j-2ci-2p+c)$

Since $P_1(n,c) = P_1(n,1/c)$, Table 3 provides $P_1(n,c)$ for $c = 1,2,3,4,10,20$ and the reciprocals for a wide range of values of n. They by utilizing formula (4) of Section II, it is possible to obtain a reasonable upper bound for $P_1(n,c)$ as n tends to infinity. To begin with, it is observed that $P_1{}^*(c)$, where $P_1{}^*(c)$ is defined as

$$P_1{}^*(c) = \lim_{n\to\infty} P_1(n,c) = \int_0^\infty \frac{c\ \exp(-cz)dz}{c\ \exp(-xc+z)+1} \ ,$$

has by Schwartz's inequality an upper bound of $\frac{1}{2}$. A better upper bound can be obtained for $c \geq 5$ and $c \leq 1/5$ by noting that

$$\frac{c\ \exp(-cx)}{c\ \exp(-xc+x)+1} \leq \frac{c\ \exp(-xc+x)}{c\ \exp(-xc+x)+1}$$

for $0 \leq x < \infty$ and $c > 0$; hence, for $c > 1$, integration yields

$$P_1{}^*(c) \leq \int_0^\infty \frac{c\ \exp(-xc+x)dx}{c\ \exp(-xc+x)+1} = \frac{\ln(c+1)}{(c-1)} \ ;$$

therefore,

$$P_1^*(c) \leq \frac{\ln(c+1)}{(c-1)} \text{ for } c > 0$$

since it is evident from formula (4) that $P_1^*(c) = P_1^*(1/c) = P_2^*(c) = P_2^*(1/c)$. Table 3 contains limiting probabilities, P^*, which were computed by numerical integration using Simpson's rule.

The result that the "rule of nearest neighbor" will have, as n tends to infinity, limiting probabilities of error of at most $\frac{1}{2}$ is particularly interesting since, as will be shown, no such general statement can be made for the linear discriminant function when the populations are characterized by exponential distributions. Considering now the linear discriminant function for the case when the populations, F and G, are assumed to have gamma distributions, a computational formula will be developed for the probabilities of misclassification. Again, it will be assumed that the samples available from each population are equal. Since this procedure consists of computing the arithmetic mean, $(\overline{X}+\overline{Y})/2$, of the sample means and assigning Z to that population whose sample mean lies on the side of $(\overline{X}+\overline{Y})/2$ as does Z itself, the error probability, P_1, is committed if and only if

$$Z > (\overline{X}+\overline{Y})/2 \text{ and } \overline{Y} > \overline{X}$$

or

$$Z < (\overline{X}+\overline{Y})/2 \text{ and } \overline{Y} < \overline{X}.$$

Thus, by the definition of P_1 it follows that

$$P_1 = P[Z > (\overline{X}+\overline{Y})/2, \overline{Y} > \overline{X}] + P[Z < (\overline{X}+\overline{Y})/2, \overline{Y} < \overline{X}].$$

For the purpose of convenience, it is desirable to define two new random variables, S and T, where $S = n\overline{X}$ and $T = n\overline{Y}$. Let the density functions of S and T be denoted by $f(s;nr,c\mu)$ and $g(t;nr,\mu)$, respectively. The probability, P_1, can now be expressed more conveniently, as

$$P_1(n,c) = P[Z > (S+T)/2n, T > S] + P[Z < (S+T)/2n, T < S]$$

$$= \int_0^\infty f(s;nr,c\mu)ds \int_s^\infty g(t;nr,\mu)dt \int_{(s+t)/2n}^\infty f(z;r,c\mu)dz$$

$$+ \int_0^\infty f(s;nr,c\mu)ds \int_0^s g(t;nr,\mu)dt \int_0^{(s+t)/2n} f(z;r,c\mu)dz$$

As in the "rule of nearest neighbor" procedure, it can easily be shown by the following change of variables, $z' = cz$, $t' = ct$, and $s' = cs$, that the relationship between P_1 and P_2 is again given by $P_1(n,c) = P_2(n,1/c)$.

Since $P_1(n,c) = P_2(n,1/c)$, it is sufficient to obtain a computation formula for $P_1(n,c)$. The methods employed to obtain this formula are now outlined. First, it is observed that $P_1(n,c)$ can be expressed as

$$P_1(n,c) = \int_0^\infty f(s;nr,c\mu)ds \int_0^s g(t;nr,\mu)dt$$

$$+ 2\int_0^\infty f(s;nr,c\mu)ds \int_s^\infty g(t;nr,\mu)dt \int_{(s+t)/2n}^\infty f(z;r,c\mu)dz$$

$$- \int_0^\infty f(s;nr,c\mu)ds \int_0^\infty g(t;nr,\mu)dt \int_{(s+t)/2n}^\infty f(z;r,c\mu)dz$$

Now by utilizing the well known integration by parts formula,

$$\frac{1}{\Gamma(n)} \int x^{n-1} \exp(-ax)dx = -\exp(-ax) \sum_{k=0}^{n-1} \frac{x^k}{\mu^{n-k}\ \Gamma(k+1)} ,$$

it can be shown that

$$P_1(n,c) = 1 - c^{nr} \sum_{k=0}^{nr-1} \frac{(nr+k-1)!}{k!(nr-1)!(c+1)^{nr+k}}$$

$$+ \frac{2c^{nr+r}}{[(nr-1)!]^2} \sum_{k=0}^{r-1} \frac{1}{(2n)^k\ c^{r-k}} \sum_{i=0}^{k} \frac{(nr+i-1)!}{i!(k-i)!} \cdot$$

$$\sum_{j=0}^{nr+i-1} \frac{(nr+k+j-i-1)!}{j![1+c/(2n)]^{nr+i-j}\ [1+c+c/n]^{nr+k+j-i}}$$

$$- \frac{c^{nr+r}}{[(nr-1)!]^2} \sum_{k=0}^{r-1} \frac{1}{(2n)^k\ c^{r-k}} \cdot$$

$$\sum_{i=0}^{k} \frac{(nr+i-1)!(nr+k-i-1)!}{i!(k-i)![1+c/(2n)]^{nr+i}\ [c+c/(2n)]^{nr+k-i}}$$

Table 4 provides a tabulation of the probabilities of misclassification, $P_1(n,c) = P_2(n,1/c)$, for r equals 1 through 20, c=1, 2, 3, 4, 5, 10, 20 and the reciprocals, and a fairly wide range of values for n.

The probabilities of misclassification for the linear discriminant function were also examined when unequal samples were available from the populations F and G for the special case when r = 1. Using techniques analogous to those described in the preceding paragraph, it is observed that for samples of size n and m from the populations described by the distributions of F and G respectively, the relationship between P_1 and P_2 is

$$P_1(j=n,i=m,c) = P_2(j=m,i=n,1/c)$$

where

$$P_1(j=n,i=m,c) = 1 - \frac{1}{[1+1/(2j)]^j\,[1+c/(2i)]^i}$$

$$- \frac{1}{[1+i/(jc)]^j} \sum_{k=0}^{i-1} \binom{j+k-1}{k} \frac{1}{[1+(jc)/i]^k}$$

$$+ \frac{2c^j}{[1+c/(2i)]^i\,[c/j+c+i/j]^j} \sum_{k=0}^{i-1} \binom{j+k-1}{k} \frac{(1+c/2)^k}{(jc+c+i)^k}$$

Although a tabulation of the error probabilities, P_1 and P_2, when the sample size is not equal, would be of some value and interest, time limitations precluded

the computation of a table which would enumerate these probabilities.

In the special case of $r = 1$, it was possible to determine the limiting probabilities of misclassification. The procedure for obtaining the limiting probabilities is briefly outlined. When $r = 1$, the distributions F and G are exponential, and P_1 can be expressed as

$$P_1(n,c) = 1/q(n,c) + \int_0^\infty f(s;n,c\mu)ds \int_0^s g(t;n,\mu)dt$$

$$-2\int_0^\infty f(s;n,c\mu)\exp[-c\mu s/(2n)]ds \int_0^s g(t;n,\mu)\exp[-c\mu t/(2n)]dt$$

which by the change of variables, $s' = c\mu(2n+1)s/(2n)$ and $t' = \mu(2n+c)t/(2n)$ for the integral appearing first in the above expression for $P_1(n,c)$ and $t' = \mu t$ and $s' = c\mu s$ for the second integral, yields

$$P_1(n,c) = 1/q(n,c) + \int_0^\infty f(s;n,1)ds \int_0^{s/c} g(t;n,1)dt$$

$$- 2/q(n,c) \int_0^\infty f(s;n,1)ds \int_0^{h(s)} g(t\ ;n,1)dt$$

where

$$h(s) = (2n+c)s/(2nc+c)$$

and

$$q(n,c) = [1+1/(2n)]^n[1+c/(2n)]^n.$$

Now, if the simple one-one transformation,

$$x = s/(s+t)$$

$$y = s + t$$

is utilized, the above expression for $P_1(n,c)$ becomes

$$P_1(n,c) = 1/q(n,c)$$

$$+ 1/\Gamma^2(n) \int_0^\infty y^{2n-1}\exp(-y)dy \int_{c/(c+1)}^1 x^{n-1}(1-x)^{n-1}dx$$

$$- 2/[q(n,c)\Gamma^2(n)] \int_0^\infty y^{2n-1}\exp(-y)dy \int_{t(n,c)}^1 x^{n-1}(1-x)^{n-1}dx,$$

which upon integrating out y, can be expressed as

$$(8). \quad P_1(n,c) = 1/q(n,c) + 1/B(n,n) \int_{c/(c+1)}^1 x^{n-1}(1-x)^{n-1}dx$$

$$-2/[q(n,c)B(n,n)] \int_{t(n,c)}^1 x^{n-1} (1-x)^{n-1} dx$$

where

$$t(n,c) = (2nc+c)/(2n+2c+2nc)$$

and

$$B(n,n) = \Gamma^2(n)/\Gamma(2n).$$

Since it is evident when $c = 1$ that $P_1(n,c) = \frac{1}{2}$ for all n, it remains only to consider the cases, $0 < c < 1$ and $c > 1$. By considering each case separately and applying Chebyshev's inequality to formula (8), the limiting probability of $P_1(n,c)$ is

$$(9).\quad P_1^*(n,c) = \lim_{n\to\infty} P_1(n,c) = \begin{cases} 1 - \exp[-(c+1)/2], & \text{if } 0<c<1 \\ \frac{1}{2}, & \text{if } c = 1 \\ \exp[-(c+1)/2], & \text{if } c > 1 \end{cases}$$

As mentioned previously, the limiting probabilities for the nonparametric discriminator, "rule of nearest neighbor," are at most $\frac{1}{2}$, but from formula (9) it is apparent that the limiting probabilities for the linear discriminant function are greater than $\frac{1}{2}$ for $[2(\ln 2)-1] < c < 1$.

TABLE 3

ERROR PROBABILITIES
RULE OF NEAREST NEIGHBOR
EXPONENTIAL POPULATIONS

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.4000	.3262	.2741	.2359	.1385	.0757
2	.5000	.4222	.3560	.3067	.2693	.1676	.0957
3	.5000	.4317	.3691	.3215	.2850	.1831	.1077
4	.5000	.4368	.3761	.3297	.2938	.1924	.1155
5	.5000	.4399	.3804	.3347	.2992	.1985	.1209
6	.5000	.4419	.3833	.3380	.3029	.2027	.1248
7	.5000	.4434	.3853	.3404	.3055	.2057	.1278
8	.5000	.4445	.3868	.3421	.3074	.2080	.1301
9	.5000	.4453	.3879	.3435	.3089	.2098	.1319
10	.5000	.4460	.3888	.3445	.3100	.2112	.1333
15	.5000	.4478	.3913	.3475	.3134	.2152	.1377
20	.5000	.4487	.3925	.3489	.3149	.2171	.1398
∞	.5000	.4507	.3954	.3524	.3188	.2217	.1447

TABLE 3

ERROR PROBABILITIES
RULE OF NEAREST NEIGHBOR
EXPONENTIAL POPULATIONS

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.5333	.5214	.5037	.4870	.4329	.3907
2	.5003	.4666	.4340	.4068	.3277	.2714
3	.4856	.4426	.4043	.3733	.2858	.2239
4	.4773	.4294	.3884	.3558	.2647	.1997
5	.4719	.4212	.3788	.3455	.2527	.1854
6	.4682	.4157	.3725	.3389	.2451	.1761
7	.4655	.4118	.3683	.3345	.2400	.1698
8	.4634	.4089	.3652	.3313	.2364	.1652
9	.4618	.4067	.3629	.3290	.2338	.1617
10	.4605	.4050	.3612	.3273	.2319	.1592
20	.4547	.3984	.3549	.3212	.2247	.1492
∞	.4507	.3954	.3524	.3188	.2217	.1447

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 1

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.4000	.3262	.2741	.2359	.1385	.0757
2	.5000	.3637	.2652	.2006	.1567	.0627	.0209
3	.5000	.3391	.2292	.1619	.1188	.0366	.0083
4	.5000	.3207	.2058	.1393	.0984	.0254	.0043
5	.5000	.3062	.1898	.1252	.0864	.0197	.0026
6	.5000	.2945	.1784	.1161	.0790	.0164	.0017
7	.5000	.2848	.1702	.1099	.0741	.0142	.0012
8	.5000	.2768	.1642	.1056	.0706	.0127	.0009
9	.5000	.2701	.1597	.1024	.0681	.0115	.0007
10	.5000	.2643	.1562	.1000	.0661	.0106	.0006
15	.5000	.2460	.1473	.0936	.0606	.0082	.0003
20	.5000	.2369	.1439	.0907	.0579	.0070	.0002
25	.5000	.2322	.1421	.0890	.0563	.0064	.0001
30	.5000	.2296	.1409	.0879	.0552	.0060	.0001
35	.5000	.2280	.1401	.0870	.0544	.0057	.0001
40	.5000	.2271	.1395	.0864	.0538	.0055	.0001
50	.5000	.2260	.1387	.0856	.0530	.0052	.0001
60	.5000	.2255	.1381	.0850	.0525	.0050	.0001
70	.5000	.2251	.1377	.0846	.0521	.0049	.0001
80	.5000	.2249	.1374	.0843	.0518	.0048	.0000
90	.5000	.2247	.1372	.0840	.0516	.0047	.0000
100	.5000	.2245	.1370	.0838	.0514	.0046	.0000
∞	.5000	.2231	.1353	.0821	.0498	.0041	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 1

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.5333	.5214	.5037	.4870	.4329	.3907
2	.5299	.5041	.4782	.4577	.4076	.3813
3	.5278	.4947	.4666	.4469	.4053	.3866
4	.5265	.4893	.4613	.4429	.4073	.3912
5	.5256	.4860	.4588	.4419	.4096	.3944
6	.5251	.4841	.4578	.4419	.4115	.3966
7	.5247	.4829	.4576	.4425	.4131	.3982
8	.5245	.4823	.4577	.4431	.4142	.3995
9	.5244	.4819	.4580	.4438	.4152	.4004
10	.5243	.4818	.4584	.4444	.4160	.4012
15	.5244	.4823	.4601	.4465	.4183	.4036
20	.5247	.4831	.4612	.4477	.4195	.4048
25	.5250	.4838	.4619	.4484	.4202	.4055
30	.5254	.4842	.4624	.4488	.4206	.4060
35	.5256	.4846	.4627	.4492	.4210	.4063
40	.5258	.4848	.4630	.4494	.4212	.4066
50	.5262	.4852	.4633	.4498	.4216	.4070
60	.5264	.4854	.4636	.4500	.4218	.4072
70	.5266	.4856	.4637	.4502	.4220	.4074
80	.5267	.4857	.4639	.4503	.4221	.4075
90	.5268	.4858	.4640	.4504	.4222	.4076
100	.5269	.4859	.4640	.4505	.4223	.4077
∞	.5277	.4833	.4647	.4512	.4231	.4084

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 2

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.3598	.2532	.1851	.1404	.0512	.0158
2	.5000	.3127	.1839	.1133	.0733	.0138	.0017
3	.5000	.2836	.1508	.0850	.0505	.0063	.0004
4	.5000	.2639	.1329	.0717	.0406	.0038	.0001
5	.5000	.2498	.1226	.0644	.0353	.0026	.0001
6	.5000	.2396	.1162	.0600	.0320	.0020	.0000
7	.5000	.2319	.1120	.0571	.0298	.0016	.0000
8	.5000	.2261	.1090	.0549	.0281	.0013	.0000
9	.5000	.2217	.1069	.0533	.0269	.0011	.0000
10	.5000	.2182	.1052	.0520	.0259	.0010	.0000
15	.5000	.2092	.1006	.0481	.0229	.0006	.0000
20	.5000	.2058	.0984	.0462	.0215	.0005	.0000
25	.5000	.2042	.0970	.0450	.0207	.0004	.0000
30	.5000	.2033	.0961	.0443	.0201	.0004	.0000
35	.5000	.2027	.0955	.0437	.0197	.0003	.0000
40	.5000	.2022	.0950	.0433	.0194	.0003	.0000
50	.5000	.2016	.0943	.0427	.0190	.0003	.0000
60	.5000	.2012	.0939	.0423	.0187	.0003	.0000
70	.5000	.2009	.0935	.0421	.0185	.0003	.0000
80	.5000	.2007	.0933	.0419	.0184	.0003	.0000
90	.5000	.2005	.0931	.0417	.0183	.0003	.0000
100	.5000	.2004	.0929	.0416	.0182	.0002	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 2

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.4999	.4487	.4071	.3770	.3109	.2807
2	.4782	.4102	.3678	.3423	.2976	.2794
3	.4658	.3946	.3566	.3354	.2979	.2806
4	.4580	.3877	.3533	.3343	.2986	.2812
5	.4527	.3846	.3526	.3344	.2991	.2815
6	.4491	.3832	.3526	.3348	.2994	.2817
7	.4466	.3827	.3528	.3351	.2997	.2819
8	.4448	.3826	.3530	.3354	.2999	.2820
9	.4435	.3826	.3533	.3356	.3000	.2821
10	.4426	.3827	.3535	.3358	.3001	.2821
15	.4409	.3833	.3541	.3363	.3004	.2823
20	.4407	.3837	.3544	.3366	.3005	.2824
25	.4409	.3840	.3546	.3367	.3006	.2824
30	.4410	.3841	.3547	.3368	.3007	.2825
35	.4412	.3842	.3548	.3369	.3007	.2825
40	.4413	.3843	.3549	.3370	.3008	.2825
50	.4415	.3845	.3550	.3371	.3008	.2825
60	.4416	.3845	.3550	.3371	.3008	.2826
70	.4417	.3846	.3551	.3371	.3008	.2826
80	.4417	.3846	.3551	.3372	.3009	.2826
90	.4418	.3847	.3552	.3372	.3009	.2826
100	.4418	.3847	.3552	.3372	.3009	.2826

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 3

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.3266	.1998	.1278	.0859	.0197	.0035
2	.5000	.2719	.1317	.0675	.0368	.0034	.0002
3	.5000	.2412	.1046	.0485	.0237	.0012	.0000
4	.5000	.2223	.0918	.0403	.0184	.0006	.0000
5	.5000	.2100	.0849	.0359	.0156	.0004	.0000
6	.5000	.2018	.0807	.0331	.0138	.0003	.0000
7	.5000	.1961	.0779	.0312	.0126	.0002	.0000
8	.5000	.1920	.0758	.0298	.0117	.0001	.0000
9	.5000	.1891	.0743	.0287	.0110	.0001	.0000
10	.5000	.1869	.0730	.0278	.0105	.0001	.0000
15	.5000	.1815	.0694	.0252	.0090	.0001	.0000
20	.5000	.1794	.0675	.0240	.0083	.0000	.0000
25	.5000	.1782	.0664	.0232	.0078	.0000	.0000
30	.5000	.1774	.0657	.0227	.0076	.0000	.0000
35	.5000	.1769	.0651	.0224	.0074	.0000	.0000
40	.5000	.1765	.0648	.0221	.0072	.0000	.0000
50	.5000	.1759	.0642	.0217	.0070	.0000	.0000
60	.5000	.1755	.0638	.0215	.0069	.0000	.0000
70	.5000	.1752	.0636	.0213	.0068	.0000	.0000
80	.5000	.1750	.0634	.0212	.0067	.0000	.0000
90	.5000	.1749	.0632	.0211	.0067	.0000	.0000
100	.5000	.1747	.0631	.0210	.0066	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 3

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.4660	.3876	.3361	.3041	.2471	.2260
2	.4326	.3443	.3004	.2770	.2374	.2199
3	.4154	.3307	.2931	.2728	.2353	.2173
4	.4056	.3260	.2913	.2718	.2341	.2158
5	.3997	.3242	.2908	.2713	.2333	.2148
6	.3960	.3236	.2905	.2710	.2328	.2141
7	.3937	.3234	.2904	.2708	.2324	.2136
8	.3923	.3233	.2903	.2707	.2321	.2132
9	.3913	.3233	.2902	.2705	.2318	.2129
10	.3908	.3233	.2902	.2704	.2316	.2126
15	.3899	.3233	.2900	.2701	.2310	.2119
20	.3900	.3233	.2899	.2699	.2307	.2115
25	.3901	.3233	.2898	.2698	.2305	.2112
30	.3902	.3233	.2898	.2698	.2303	.2111
35	.3903	.3233	.2897	.2697	.2302	.2109
40	.3903	.3233	.2897	.2697	.2302	.2108
50	.3904	.3233	.2897	.2696	.2301	.2107
60	.3904	.3233	.2897	.2696	.2300	.2106
70	.3905	.3233	.2896	.2695	.2299	.2106
80	.3905	.3233	.2896	.2695	.2299	.2105
90	.3905	.3233	.2896	.2695	.2299	.2105
100	.3905	.3233	.2896	.2695	.2298	.2105

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 4

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.2974	.1588	.0892	.0533	.0078	.0008
2	.5000	.2379	.0964	.0417	.0193	.0009	.0000
3	.5000	.2076	.0750	.0289	.0117	.0003	.0000
4	.5000	.1904	.0656	.0235	.0087	.0001	.0000
5	.5000	.1801	.0606	.0206	.0071	.0001	.0000
6	.5000	.1736	.0574	.0187	.0061	.0000	.0000
7	.5000	.1693	.0552	.0174	.0055	.0000	.0000
8	.5000	.1663	.0536	.0165	.0050	.0000	.0000
9	.5000	.1642	.0524	.0158	.0046	.0000	.0000
10	.5000	.1626	.0514	.0152	.0044	.0000	.0000
15	.5000	.1586	.0484	.0135	.0036	.0000	.0000
20	.5000	.1567	.0469	.0127	.0032	.0000	.0000
25	.5000	.1556	.0460	.0122	.0030	.0000	.0000
30	.5000	.1549	.0454	.0119	.0029	.0000	.0000
35	.5000	.1544	.0449	.0117	.0028	.0000	.0000
40	.5000	.1540	.0446	.0115	.0027	.0000	.0000
50	.5000	.1534	.0442	.0113	.0027	.0000	.0000
60	.5000	.1531	.0439	.0111	.0026	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 4

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.4345	.3383	.2844	.2543	.2062	.1885
2	.3937	.2967	.2547	.2330	.1951	.1777
3	.3748	.2861	.2493	.2290	.1910	.1730
4	.3650	.2828	.2475	.2273	.1887	.1704
5	.3597	.2815	.2465	.2262	.1872	.1688
6	.3568	.2809	.2459	.2254	.1862	.1676
7	.3551	.2806	.2454	.2249	.1855	.1668
8	.3541	.2803	.2451	.2245	.1849	.1661
9	.3536	.2802	.2448	.2241	.1845	.1656
10	.3532	.2800	.2446	.2239	.1841	.1652
15	.3528	.2795	.2439	.2230	.1830	.1640
20	.3528	.2793	.2435	.2226	.1824	.1633
25	.3528	.2792	.2433	.2223	.1821	.1630
30	.3528	.2791	.2432	.2222	.1818	.1627
35	.3528	.2790	.2431	.2220	.1817	.1625
40	.3528	.2789	.2430	.2219	.1815	.1624
50	.3528	.2789	.2429	.2218	.1814	.1622
60	.3528	.2788	.2428	.2217	.1812	.1620

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 5

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.2714	.1269	.0629	.0335	.0031	.0002
2	.5000	.2093	.0718	.0264	.0105	.0002	.0000
5	.5000	.1565	.0439	.0120	.0033	.0000	.0000
10	.5000	.1426	.0365	.0084	.0018	.0000	.0000

R = 6

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.2480	.1019	.0446	.0213	.0013	.0000
2	.5000	.1850	.0542	.0170	.0058	.0001	.0000
5	.5000	.1374	.0321	.0071	.0015	.0000	.0000
10	.5000	.1256	.0261	.0047	.0008	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 5

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.4057	.2984	.2455	.2181	.1759	.1596
2	.3605	.2607	.2208	.2000	.1627	.1458
5	.3283	.2482	.2121	.1914	.1527	.1349
10	.3236	.2459	.2093	.1882	.1488	.1307
50	.3227	.2440				

R = 6

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.3794	.2658	.2153	.1904	.1519	.1363
2	.3320	.2321	.1938	.1735	.1371	.1208
5	.3026	.2209	.1843	.1637	.1259	.1090
10	.2988	.2180	.1808	.1599	.1217	.1045

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 7

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.2268	.0821	.0319	.0136	.0005	.0000
2	.5000	.1642	.0414	.0111	.0032	.0000	.0000
5	.5000	.1214	.0237	.0042	.0007	.0000	.0000
10	.5000	.1111	.0188	.0026	.0003	.0000	.0000

R = 8

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.2077	.0664	.0230	.0088	.0002	.0000
2	.5000	.1463	.0319	.0074	.0018	.0000	.0000
5	.5000	.1078	.0175	.0025	.0003	.0000	.0000
10	.5000	.0985	.0136	.0015	.0001	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 7

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.3555	.2387	.1911	.1682	.1322	.1172
2	.3073	.2087	.1715	.1517	.1163	.1008
5	.2809	.1979	.1612	.1409	.1046	.0888
10	.2775	.1945	.1574	.1369	.1003	.0843

R = 8

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.3336	.2159	.1711	.1497	.1155	.1012
2	.2857	.1890	.1526	.1333	.0992	.0846
5	.2621	.1781	.1417	.1220	.0874	.0727
10	.2588	.1744	.1377	.1178	.0831	.0684

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 9

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1904	.0540	.0166	.0057	.0001	.0000
2	.5000	.1308	.0247	.0049	.0010	.0000	.0000
5	.5000	.0961	.0130	.0015	.0002	.0000	.0000
10	.5000	.0875	.0099	.0009	.0001	.0000	.0000

R = 10

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1747	.0440	.0121	.0037	.0000	.0000
2	.5000	.1174	.0193	.0033	.0006	.0000	.0000
5	.5000	.0859	.0097	.0009	.0001	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 9

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.3137	.1965	.1542	.1341	.1014	.0877
2	.2667	.1720	.1364	.1175	.0850	.0713
5	.2454	.1609	.1251	.1060	.0734	.0599
10	.2421					

R = 10

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2954	.1798	.1397	.1206	.0893	.0762
2	.2498	.1571	.1223	.1040	.0730	.0603
5	.2305	.1458	.1108	.0925	.0619	.0495
10	.2270	.1417	.1066	.0883	.0579	.0457

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 11

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1605	.0359	.0088	.0025	.0000	.0000
2	.5000	.1056	.0151	.0022	.0003	.0000	.0000
5	.5000	.0770	.0073	.0005	.0000	.0000	.0000
10	.5000	.0694	.0052	.0003	.0000	.0000	.0000

R = 12

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1475	.0294	.0065	.0016	.0000	.0000
2	.5000	.0953	.0119	.0015	.0002	.0000	.0000
5	.5000	.0691	.0054	.0003	.0000	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 11

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2787	.1652	.1270	.1087	.0788	.0664
2	.2348	.1439	.1099	.0923	.0630	.0511
5	.2170	.1324	.0984	.0809	.0523	.0410

R = 12

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2633	.1523	.1159	.0983	.0696	.0580
2	.2212	.1321	.0990	.0821	.0544	.0435
5	.2046	.1205	.0876	.0710	.1443	.0341
10	.2009	.1163	.0835	.0670	.0408	.0309

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 13

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1358	.0242	.0048	.0011	.0000	.0000
2	.5000	.0861	.0094	.0010	.0001	.0000	.0000
5	.5000	.0621	.0041	.0002	.0000	.0000	.0000
10	.5000	.0554	.0028	.0001	.0000	.0000	.0000

R = 14

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1250	.0199	.0035	.0007	.0000	.0000
2	.5000	.0781	.0074	.0007	.0001	.0000	.0000
10	.5000	.0496	.0021	.0001	.0000	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 13

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2492	.1409	.1060	.0891	.0617	.0508
2	.2089	.1215	.0894	.0731	.0471	.0371

R = 14

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2361	.1307	.0971	.0808	.0547	.0445
2	.1977	.1119	.0808	.0653	.0408	.0316
5	.1829	.1004	.0698	.0549	.0320	.0237
10	.1789	.0962				

TABLE 4

LINEAR CISCRIMIVANT FUNCTION
ERRCR PROBABILITIES
FCR GAMMA POPULATIONS

R = 15

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1152	.0165	.0026	.0005	.0000	.0000
2	.5000	.0709	.0059	.0005	.0000	.0000	.0000
0	.5000	.0504	.0023	.0001	.0000	.0000	.0000

R = 16

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.1063	.0136	.0020	.0003	.0000	.0000
2	.5000	.0645	.0047	.0003	.0000	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 15

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2241	.1214	.0891	.0734	.0486	.0391
2	.1875	.1033	.0731	.0583	.0355	.0271

R = 16

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2129	.1131	.0819	.0668	.0433	.0343
2	.1780	.0954	.0663	.0522	.0309	.0232

TABLE 4

LINEAR CISCRIMINANT FUNCTION
ERRCR PROBABILITIES
FOR GAMMA POPULATIONS

R = 17

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.0981	.0113	.0015	.0002	.0000	.0000
2	.5000	.0587	.0037	.0002	.0000	.0000	.0000

R = 18

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.0906	.0094	.0011	.0001	.0000	.0000
2	.5000	.0536	.0029	.0001	.0000	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 17

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.2026	.1054	.0753	.0609	.0385	.0302
2	.1693	.0882	.0601	.0468	.0269	.0199

R = 18

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.1929	.0985	.0693	.0555	.0344	.0266
2	.1612	.0816	.0546	.0419	.0234	.0171

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 19

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.0838	.0078	.0008	.0001	.0000	.0000
2	.5000	.0489	.0023	.0001	.0000	.0000	.0000

R = 20

N\C	1.0	2.0	3.0	4.0	5.0	10.0	20.0
1	.5000	.0775	.0065	.0006	.0001	.0000	.0000
2	.5000	.0448	.0019	.0001	.0000	.0000	.0000

TABLE 4

LINEAR DISCRIMINANT FUNCTION
ERROR PROBABILITIES
FOR GAMMA POPULATIONS

R = 19

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.1840	.0920	.0639	.0506	.0307	.0235
2	.1536	.0756	.0496	.0376	.0205	.0147

R = 20

N\C	.5000	.3333	.2500	.2000	.1000	.0500
1	.1756	.0861	.0590	.0462	.0274	.0207
2	.1465	.0701	.0452	.0338	.0179	.0126

SECTION IV

SUMMARY AND CONCLUSION

Section II of this paper briefly summarizes some of the work accomplished by Hodges and Fix in [3]. Their investigation was concerned with the computation of the probabilities of misclassification for various nonparametric procedures assuming some parametric form of the distribution which describes the populations. The error probabilities for the "optimum" parametric procedure were also computed and compared with the nonparametric error probabilities. The investigation considered the two population classification problem when the populations have normal distributions with equal covariance matrices. The parametric procedure employed was the linear discriminant function which is the appropriate method in this situation, and the primary nonparametric procedure considered was the "rule of the nearest neighbor." The above two procedures were compared by computing the probabilities of misclassification. The results of this investigation indicated that the "rule of nearest neighbor" gave "reasonable" error probabilities.

Section III also considers the two population classification problem, but the investigation is primarily concerned with the performance of the linear discriminant

function if the actual densities which describe the populations are not normal, but in fact gamma with density functions defined by formulas (5) and (6) of Section III. Also included in Section III is a limited investigation of the "rule of nearest neighbor" when the populations are assumed to be exponential. Evaluation of the performance of both the linear discriminant function and the "rule of nearest neighbor" was accomplished by computation of the probabilities of misclassification.

When the population densities are assumed to be exponential, Table 3 and Table 4, for the case $r = 1$, provide a means of comparing the performance of the linear discriminant function and the "rule of nearest neighbor." An examination of these tables indicates that both procedures can result in "high" probabilities of error, particularly when c assumes values near one, since for small sample sizes, both procedures can result in error probabilities which are greater than $\frac{1}{2}$. Although even as n, the sample size from each population, tends to infinity, the linear discriminant function has error probabilities greater than $\frac{1}{2}$ for $[2(\ln 2)-1] < c < 1$, it is of interest to note that "the rule of nearest neighbor" in this situation will always have error probabilities less than or equal to $\frac{1}{2}$. Also, depending upon the importance of each type of

error, it is possible for the linear discriminant function to be a "fairly useful" procedure since one error probability is usually "small." Table 4 also shows that as r increases, the probabilities of misclassification decrease. This result was anticipated since for increasing r, the gamma distribution approaches the normal distribution by the Central Limit Theorem.

The following recommendations are made on the basis of this paper.

(i) Investigate the performance of the nonparametric procedure, using k = 3 instead of the "rule of nearest neighbor," k = 1.

(ii) Investigate the performance of the nonparametric procedures proposed by Hodges and Fix in [2] employing different distance functions.

(iii) Develop a more satisfactory computational formula for the linear discriminant function when the populations are assumed to be gamma in the situation when r and n are large since the formula used in this paper required many hours of computer time.

(iv) Investigate the performance of the linear discriminant function and other nonparametric procedures for other distributions. A cursory investigation was made for the beta distribution and the analysis appears to be more difficult.

(v) Compare the performance of Bayesian parametric and nonparametric classification procedures.

(vi) Investigate the classification problem when there are more than two populations.

BIBLIOGRAPHY

1. Eaton, Thomas E. Jr., "Discrimination Procedures, Small Sample Performance," Thesis at United States Naval Postgraduate School, 1963.

2. Fix, Evelyn and Hodges, J. L., Jr., "Nonparametric Discrimination. I. Consistency Properties," Technical Report prepared under contract between University of California and School of Aviation Medicine, Randolph Field, Texas, 22 pp.

3. Fix, Evelyn and Hodges, J. L., Jr., "Nonparametric Discrimination: Small Sample Performance," Technical Report prepared under contract between University of California and School of Aviation Medicine, Randolph Field, Texas.

4. Hodges, J. L., Jr., "Discriminatory Analysis. I. Survey of Discriminatory Analysis," Technical Report prepared under contract between University of California and School of Aviation Medicine, Randolph Field, Texas.